Symmetrization approach to concentration inequalities for empirical processes.

Dmitry Panchenko
Massachusetts Institute of Technology

Abstract

We introduce a symmetrization technique that allows us to translate a problem of controlling the deviation of some functionals on a product space from their mean into a problem of controlling the deviation between two independent copies of the functional. As an application we give a new easy proof of Talagrand's concentration inequality for empirical processes, where besides symmetrization we use only Talagrand's concentration inequality on the discrete cube $\{-1,+1\}^n$. As another application of this technique we prove new Vapnik-Chervonenkis type inequalities. For example, for VC-classes of functions we prove a classical inequality of Vapnik and Chervonenkis only with normalization by the sum of variance and sample variance.

1991 AMS subject classification: primary 62G05, secondary 62G20, 60F15
Keywords and phrases: empirical processes, concentration inequalities

1 Introduction and main results.

Let us consider a measurable space Ω with probability measure μ, and the corresponding product space (Ω^n, μ^n). Given a class of measurable functions $\mathcal{F} = \{f : \Omega \to \mathbb{R}\}$, we consider a functional

$$Z(x) = \sup_{\mathcal{F}} \sum_{i=1}^{n} f(x_i)$$

where $x = (x_1, \ldots, x_n) \in \Omega^n$, which is usually called an empirical process. To avoid measurability problems we will assume that $\mathcal{F}$ is countable, or even finite. Our main interest is to study the deviation inequalities for this (or similar) functional from its mean. The main observation of this paper is that this problem can be translated into a problem of studying $Z(x) - Z(y)$, where y lives on a separate copy of Ω^n. This new problem turns out to be easier, at least in the examples we have in mind here, as it can be handled with Talagrand's convex distance inequality on $\{-1, +1\}^n$ which is the simplest case of convex distance inequality (see Talagrand (1995)).

As a first example of application of this technique we will give an easy proof of Talagrand's concentration inequality for $Z(x)$. As a second example, we will prove new Vapnik-Chervonenkis type inequalities.

Let us start by proving the main result that will allow us to implement the mentioned symmetrization. For $x \in \mathbb{R}$ we will denote $(x)_+ = \max(x, 0)$.

Lemma 1 *If ξ and ν are r.v.s such that for any number $a \in \mathbb{R}$ and a function $\phi(x) = (x-a)_+$*

$$\mathbb{E}\phi(\xi) \le \mathbb{E}\phi(\nu)$$

and for some $\Gamma \ge 1, \gamma > 0$ and for all $t \ge 0$

$$\mathbb{P}(\nu \ge t) \le \Gamma e^{-\gamma t},$$

then for all $t \ge 0$

$$\mathbb{P}(\xi \ge t) \le \Gamma e^{1-\gamma t}.$$

Proof. Let $\phi(x) = (x - a)_+$ for some $a \in \mathbb{R}$ that will be chosen later. Note that ϕ is nondecreasing. For $t > 0$ we can write

$$\mathbb{P}(\xi \ge t) \le \frac{\mathbb{E}\phi(\xi)}{\phi(t)} \le \frac{\mathbb{E}\phi(\nu)}{\phi(t)} = \frac{1}{\phi(t)}\Big(\phi(0) + \int_0^\infty \phi'(x)\mathbb{P}(\nu \ge x)dx\Big)$$
$$\le \frac{1}{\phi(t)}\Big(\phi(0) + \Gamma \int_0^\infty \phi'(x)e^{-\gamma x}dx\Big),$$

where we used integration by parts. Since $\Gamma \ge 1$, we can assume that $t \ge \gamma^{-1}$. Take

$$a = t - \frac{1}{\gamma}, \quad \phi(x) = \Big(x - t + \frac{1}{\gamma}\Big)_+.$$

Then $\phi(t) = \gamma^{-1}$, $\phi(0) = 0$ and

$$\int_0^\infty \phi'(x)e^{-\gamma x}dx = \int_{t-\gamma^{-1}}^\infty e^{-\gamma x}dx = \gamma^{-1}e^{1-\gamma t},$$

which gives $\mathbb{P}(\xi \geq t) \leq \Gamma e^{1-\gamma t}$.

□

It is clear that the Lemma can be stated in more generality, for instance, we could consider the case of tails $\Gamma e^{-\gamma t^\alpha}$ for $\alpha > 0$. But it is irrelevant for the applications of this paper. The main consequence is given by the following corollary.

Corollary 1 *Let $\xi_i(x, y) : \Omega^n \times \Omega^n \to R$, $1 \leq i \leq 3$ be measurable functions defined on two copies of Ω^n and let*

$$\xi_i'(x) = \int_{\Omega^n} \xi_i(x, y)d\mu^n(y).$$

If $\xi_3 \geq 0$ and for all $t \geq 0$

$$\mu^{2n}(\xi_1 \geq \xi_2 + (\xi_3 t)^{1/2}) \leq \Gamma e^{-\gamma t},$$

then for all $t \geq 0$

$$\mu^n(\xi_1' \geq \xi_2' + (\xi_3' t)^{1/2}) \leq \Gamma e^{1-\gamma t}.$$

Proof. Since $\sqrt{ab} = \inf_{\delta>0}(\delta a + b/(4\delta))$ we can rewrite the events

$$\Big\{\xi_1 \geq \xi_2 + (\xi_3 t)^{1/2}\Big\} = \Big\{\sup_{\delta>0} 4\delta(\xi_1 - \xi_2 - \delta\xi_3) \geq t\Big\}$$

and, similarly,

$$\Big\{\xi_1' \geq \xi_2' + (\xi_3' t)^{1/2}\Big\} = \Big\{\sup_{\delta>0} 4\delta(\xi_1' - \xi_2' - \delta\xi_3') \geq t\Big\}.$$

Let us denote

$$\xi = \sup_{\delta>0} 4\delta(\xi_1 - \xi_2 - \delta\xi_3), \quad \nu = \sup_{\delta>0} 4\delta(\xi_1' - \xi_2' - \delta\xi_3').$$

Clearly,

$$\nu = \sup_{\delta>0} \int 4\delta(\xi_1 - \xi_2 - \delta\xi_3)d\mu^n(y) \leq \int \xi d\mu^n(y),$$

and, thus, by Jensen's inequality, for any nondecreasing convex funcion ϕ

$$\int \phi(\nu)d\mu^n(x) \leq \int \phi\Big(\int \xi d\mu^n(y)\Big)d\mu^n(x) \leq \int \phi(\xi)d\mu^n(x)d\mu^n(y).$$

Lemma 1 implies the result.

□

As we mentioned above, besides the symmetrization of Corollary 1 we will need Talagrand's convex distance inequality, which we will formulate now.

Consider the space $\{0,1\}^n$ with uniform measure $\mathbb{P}_\varepsilon$. If $\varepsilon \in \{0,1\}^n$ and $\mathcal{A} \subseteq \{0,1\}^n$, denote

$$U_{\mathcal{A}}(\varepsilon) = \{(s_i)_{i\le n} \in \{0,1\}^n, \exists \varepsilon' \in \mathcal{A}, s_i = 0 \Rightarrow \varepsilon'_i = \varepsilon_i\}.$$

Denote the "convex hull" distance between the point ε and a set $\mathcal{A}$ as

$$f_c(\mathcal{A},\varepsilon) = \inf\{|s| : s \in \mathrm{conv}U_{\mathcal{A}}(\varepsilon)\},$$

where $|s|$ denotes the Euclidean norm of s. The concentration inequality of Talagrand (Theorem 4.3.1 in [14]) states the following.

Proposition 1 *For any $\alpha \ge 0$*

$$\mathbb{P}_\varepsilon(f_c^2(\mathcal{A},\varepsilon) \ge t) \le \frac{1}{\mathbb{P}_\varepsilon(\mathcal{A})^\alpha} \exp\Big\{-\frac{\alpha}{\alpha+1}t\Big\}. \tag{1.1}$$

Remark. In [14] this result was formulated for $\alpha \ge 1$, but it was proven (and used) for $\alpha \ge 0$.

The main feature of this distance is that if $f_c^2(\mathcal{A},\varepsilon) \le t$, then (Theorem 4.1.2 in [14])

$$\forall(\lambda_i)_{i\le n} \quad \exists \varepsilon' \in \mathcal{A} \qquad \sum_{i=1}^n \lambda_i I(\varepsilon'_i \ne \varepsilon_i) \le (t\sum_{i=1}^n \lambda_i^2)^{1/2}. \tag{1.2}$$

We will start by giving a new proof of Talagrand's concentration inequality for empirical processes.

2 Talagrand's concentration inequality for empirical processes.

For simplicity of notations from now on we will write $\mathbb{P}$ to denote any probability measure, and $\mathbb{P}_\xi$ to specify the distribution on the space of random variable ξ, with all other variables fixed. Similarly, to denote the expectation we will write $\mathbb{E}$ and $\mathbb{E}_\xi$.

Let us define a mixed uniform variance as

$$V = \mathbb{E}_y \sup_{f\in\mathcal{F}} \sum_{i=1}^n (f(x_i) - f(y_i))^2. \tag{2.1}$$

In a sense, V is a uniform version of the sum of variance and sample variance, since in the case when $\mathcal{F}$ consists of one function, this is exactly what it is. Clearly, V is a function of x. The following theorem holds.

Theorem 1 *Let V be defined by (2.1). Then for any $\alpha > 0$*

$$\mathbb{P}\Big(\sup_{f\in\mathcal{F}} \sum_{i=1}^n f(x_i) \ge \mathbb{E}\sup_{f\in\mathcal{F}} \sum_{i=1}^n f(x_i) + 2\sqrt{Vt}\Big) \le 2^{\alpha+1}\exp\Big\{1 - \frac{\alpha}{\alpha+1}t\Big\}$$

and

$$\mathbb{P}\Big(\sup_{f\in\mathcal{F}}\sum_{i=1}^n f(x_i) \le \mathbb{E}\sup_{f\in\mathcal{F}}\sum_{i=1}^n f(x_i) - 2\sqrt{Vt}\Big) \le 2^{\alpha+1}\exp\Big\{1-\frac{\alpha}{\alpha+1}t\Big\}$$

Remark. One can optimize the bound over α, which would give that for $t \ge \log 2$, the bound can be written as $2\exp\{1-(\sqrt{t}-\sqrt{\log 2})^2\}$.

Proof. We will only prove the upper tail, since the proof of the lower tail is exactly the same, once one switches Z and EZ. Since

$$\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{i=1}^n f(x_i) = \mathbb{E}_y\sup_{f\in\mathcal{F}}\sum_{i=1}^n f(y_i)$$

Corollary 1 implies that it is enough to prove that

$$\mathbb{P}\Big(\sup_{f\in\mathcal{F}}\sum_{i=1}^n f(x_i) \ge \sup_{f\in\mathcal{F}}\sum_{i=1}^n f(y_i) + 2\sqrt{Wt}\Big) \le 2^{\alpha+1}\exp\Big\{-\frac{\alpha}{\alpha+1}t\Big\},$$

where $W = \sup_{f\in\mathcal{F}}\sum_{i=1}^n (f(x_i)-f(y_i))^2$. For any $(x_1,\dots,x_n,y_1,\dots,y_n)$, let Π be the set of permutations of these coordinates such that, for each $1\le i\le n$, $\pi(x_i),\pi(y_i)\in\{x_i,y_i\}$, and let $\mathbb{P}_\pi$ denote the uniform probability measure on Π. Since the above probability is invariant with respect to any $\pi\in\Pi$, it is enough to show that for any fixed $x=(x_1,\dots,x_n)$ and $y=(y_1,\dots,y_n)$ the probability over permutations

$$\mathbb{P}_\pi\Big(\sup_{f\in\mathcal{F}}\sum_{i=1}^n f(z_i^1) \ge \sup_{f\in\mathcal{F}}\sum_{i=1}^n f(z_i^2) + 2\sqrt{Wt}\Big) \le 2^{\alpha+1}\exp\Big\{-\frac{\alpha}{\alpha+1}t\Big\},$$

where $z_i^1=\pi(x_i)$ and $z_i^2=\pi(y_i)$. Note that W is invariant under permutations. We can rewrite it differently in terms of an i.i.d. Bernoulli sequence $\varepsilon=(\varepsilon_1\dots,\varepsilon_n)$, i.e. $\mathbb{P}(\varepsilon_i=0)=\mathbb{P}(\varepsilon_i=1)=1/2$. Namely, we can write

$$f(z_i^1)=f(y_i)+\varepsilon_i(f(x_i)-f(y_i)),\ \ f(z_i^2)=f(x_i)-\varepsilon_i(f(x_i)-f(y_i)),$$

and instead of permutations look at the distribution $\mathbb{P}_\varepsilon$ of ε. For any $f\in\mathcal{F}$ let us denote $c_f=\sum f(y_i)$, $c_f'=\sum f(x_i)$, and $f_i=(f(x_i)-f(y_i))$. Then, we need to prove that

$$\mathbb{P}_\varepsilon\Big(\sup_{f\in\mathcal{F}}\Big(c_f+\sum_{i=1}^n\varepsilon_i f_i\Big) \ge \sup_{f\in\mathcal{F}}\Big(c_f'-\sum_{i=1}^n\varepsilon_i f_i\Big) + 2\Big(t\ \sup_{f\in\mathcal{F}}\sum_{i=1}^n f_i^2\Big)^{1/2}\Big) \le 2^{\alpha+1}\exp\Big\{-\frac{\alpha}{\alpha+1}t\Big\}.$$

But this is an easy consequence of Proposition 1. Let us consider the functionals

$$\Phi(\varepsilon)=\sup_{f\in\mathcal{F}}\Big(c_f+\sum_{i=1}^n\varepsilon_i f_i\Big),\ \ \Phi'(\varepsilon)=\sup_{f\in\mathcal{F}}\Big(c_f'-\sum_{i=1}^n\varepsilon_i f_i\Big).$$

They are both convex, with the Lipschitz norm bounded by

$$\|\Phi\|_L, \|\Phi'\|_L \le \Big(\sup_{f\in\mathcal{F}}\sum f_i^2\Big)^{1/2}.$$

Also, by symmetry, they have the same median, $M = M(\Phi) = M(\Phi')$ with respect to $\mathbb{P}_\varepsilon$. We will now show that from the convexity of Φ and Φ' and Proposition 1 it follows

$$\mathbb{P}_\varepsilon(\Phi(\varepsilon) \geq M + \|\Phi\|_L\sqrt{t}) \leq 2^\alpha \exp\Big\{-\frac{\alpha}{\alpha+1}t\Big\}, \tag{2.2}$$

and

$$\mathbb{P}_\varepsilon(\Phi'(\varepsilon) \leq M - \|\Phi'\|_L\sqrt{t}) \leq 2^\alpha \exp\Big\{-\frac{\alpha}{\alpha+1}t\Big\}. \tag{2.3}$$

Let us recall how this is usually done (see Ledoux and Talagrand (1991)). If we consider the set $\mathcal{A} = \{\varepsilon : \Phi(\varepsilon) \leq M\}$, then $\mathrm{P}(\mathcal{A}) \geq 1/2$ and by convexity of Φ, $\mathrm{conv}\mathcal{A} = \mathcal{A}$. This, together with the Lipschitz condition, implies that

$$\{f_c^2(\mathcal{A}, \varepsilon) \leq t\} \subseteq \{\Phi(\varepsilon) \leq M + \|\Phi\|_L\sqrt{t}\}.$$

Thus, the right tail (2.2) follows from Proposition 1. Similarly, if we consider the set

$$\mathcal{B} = \{\varepsilon : \Phi'(\varepsilon) \leq M - \|\Phi'\|_L\sqrt{t}\},$$

then

$$\{f_c^2(\mathcal{B}, \varepsilon) \leq t\} \subseteq \{\Phi'(\varepsilon) \leq M\}.$$

By Proposition 1,

$$\frac{1}{2} \leq \mathbb{P}(f_c^2(\mathcal{B}, \varepsilon) \geq t) \leq \frac{1}{\mathbb{P}(\mathcal{B})^\alpha} \exp\Big\{-\frac{\alpha}{\alpha+1}t\Big\}.$$

We can rewrite this as

$$\mathbb{P}(\mathcal{B}) \leq 2^\beta \exp\Big\{-\frac{\beta}{\beta+1}t\Big\},$$

where $\beta = 1/\alpha$. But since α is arbitrary, this proves the lower tail (2.3), which completes the proof of the theorem.

□

This result is an intermediate step in obtaining the concentration inequality for $Z(x)$ in its final form, since V still depends on x. Notice that here we did not assume any boundedness of $f \in \mathcal{F}$, and the result is of somewhat similar nature as the self-normalization phenomenon in the one-dimensional case (see Giné et. al. (1997), or Shao(1997)). Under the additional assumption that $f \in \mathcal{F}$ are uniformly bounded one can proceed by controlling the deviation of V (or W) from its expectation, which is done in a usual way, either via control by two points as in Talagrand (1996) plus some truncation argument, or via a sharp concentration inequality of Boucheron et. al. (2000).

Let us assume now that

$$\forall f \in \mathcal{F}\ \forall x \in \Omega,\ \ -\frac{1}{2} \leq f(x) \leq \frac{1}{2}.$$

If we introduce $V_i = \mathbb{E}_y \sup_{f\in\mathcal{F}} \sum_{j\neq i}(f(x_j) - f(y_j))^2$ then, it is easy to see that

$$0 \leq V - V_i \leq 1 \quad \text{and} \quad \sum_{i=1}^n (V - V_i) \leq V.$$

Under these conditions, Theorem 6 in Boucheron et. al. (2000) states that for all $t \geq 0$,

$$\mathbb{P}\Big(V \geq \mathbb{E}V + t\Big) \leq \exp\Big\{-\mathbb{E}V h\Big(\frac{t}{\mathbb{E}V}\Big)\Big\}, \tag{2.4}$$

where $h(x) = (1+x)\log(1+x) - x$. Since $h(x) \geq x^2/(2 + 2x/3)$, (2.4) implies Bernstein's inequality

$$\mathbb{P}\Big(V \geq \mathbb{E}V + t\Big) \leq \exp\Big\{-\frac{t^2}{2\mathbb{E}V + 2t/3}\Big\},$$

which can be equivalently written as

$$\mathbb{P}\Big(V \geq \mathbb{E}V + \frac{1}{3}(18\mathbb{E}Vt + t^2)^{1/2} + \frac{t}{3}\Big) \leq e^{-t}.$$

More generally, if $-b \leq f(x) \leq b$, then

$$\mathbb{P}\Big(V \geq \mathbb{E}V + \frac{2b}{3}(18\mathbb{E}Vt + 4b^2t^2)^{1/2} + \frac{4b^2t}{3}\Big) \leq e^{-t}.$$

Combining this with Theorem 1 we get the following corollary.

Corollary 2 *If $-b \leq f(x) \leq b$ then for all $t \geq \log 2$,*

$$\mathbb{P}\Big(|Z - \mathbb{E}Z| \geq 2\Big(t\Big(\mathbb{E}V + \frac{2b}{3}(18\mathbb{E}Vt + 4b^2t^2)^{1/2} + \frac{4b^2t}{3}\Big)\Big)^{1/2}\Big) \leq 4e^{1-(\sqrt{t}-\sqrt{\log 2})^2} + e^{-t}. \tag{2.5}$$

□

It is clear, that in the range of parameters $1 \ll t \ll \mathbb{E}V/b^2$, the bound of the Corollary will be dominated by the term $\sim 2\sqrt{\mathbb{E}Vt}$. For this range, it improves upon the control of the lower tail given by Theorem 12 in Massart (2000), which states

$$\mathbb{P}\Big(Z \leq \mathbb{E}Z - 2\sqrt{1.35\mathbb{E}Vt} - 3.5bt\Big) \leq e^{-t}. \tag{2.6}$$

Actually, one can check that

$$2\Big(t\Big(\mathbb{E}V + \frac{2b}{3}(18\mathbb{E}Vt + 4b^2t^2)^{1/2} + \frac{4b^2t}{3}\Big)\Big)^{1/2} \leq 2\sqrt{1.35\mathbb{E}Vt} + 3.5bt$$

for all parameters $b, \mathbb{E}V, t$. Unfortunately, (2.5) and (2.6) are not comparable in all range of parameters, mainly, because of the term $\exp\{-(\sqrt{t} - \sqrt{\log 2})^2\}$.

Finally, for more results in this

3 Vapnik-Chervonenkis type inequalities.

In this section we are trying to control the functional $Q_n f$ uniformly over the class $\mathcal{F}$, where

$$Q_n f = Pf - P_n f \quad \text{or} \quad Q_n f = P_n f - Pf$$

and

$$Pf = \int f(x) dP(x), \;\; P_n f = \frac{1}{n} \sum_{i=1}^{n} f(x_i).$$

The difference from the previous section is that now the bounds on $Q_n f$ will depend on f and will reflect that the function f with a smaller variance should have a tighter bound. The results of this section are in a spirit of Vapnik and Chervonenkis (1968) and Panchenko (2002).

Corresponding to $Q_n f$, let us introduce

$$S_n f = \frac{1}{n} \sum_{i=1}^{n} (f(y_i) - f(x_i)) \quad \text{or} \quad S_n f = \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - f(y_i)).$$

Finally, we define

$$R_n f = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i (f(y_i) - f(x_i)),$$

$$Wf = W(f, x, y) = \frac{4}{n} \sum_{i=1}^{n} (f(y_i) - f(x_i))^2, \quad Vf = V(f, x) = \mathbb{E}_y W(f, x, y).$$

As one of the consequences of our approach we will give a uniform control of $Q_n f / (Vf)^{1/2}$ for VC-subgraph classes of functions. The original result of Vapnik and Chervonenkis [17] provided a uniform control for $Q_n f / (Pf)^{1/2}$ for VC-classes of functions taking values $f \in \{0, 1\}$ (and a simple generalization for VC-major classes taking values in $[0, 1]$). The fact that we can substitute Pf by Vf gives a new way to control $Q_n f$.

Let us introduce a function $\Phi(f, x, y)$ which is invariant over all permutations of (x, y) that switch only the same coordinates of x and y. Assume that for some fixed $\beta \in (0, 1)$ and for any fixed (x, y) we have

$$\mathbb{P}_\varepsilon \Big(\sup_{f \in \mathcal{F}} (R_n f - \Phi(f, x, y)) > 0 \Big) < 1 - \beta. \tag{3.1}$$

Then the following theorem holds.

Theorem 2 *Assume that (3.1) holds. Then for any $t \geq \log \beta^{-1}$,*

$$\mathbb{P}\Big(\exists f \in \mathcal{F} \; Q_n f \geq \mathbb{E}_y \Phi(f, x, y) + \sqrt{\frac{Vt}{n}} \Big) \leq \exp(1 - (\sqrt{t} - \sqrt{\log \beta^{-1}})^2).$$

Proof. We will first prove that for any $\alpha \geq 0$ the statement of the theorem holds with the right hand side substituted by $\beta^{-\alpha}\exp(1-\alpha t/(\alpha+1))$. The result will follow by optimization over α. First of all, by Corollary 1 it is enough to prove that

$$\mathbb{P}\Big(\exists f \ S_n f \geq \Phi(f,x,y) + \sqrt{\frac{Wt}{n}}\Big) \leq \frac{1}{\beta^\alpha}\exp\Big(-\frac{\alpha}{\alpha+1}t\Big).$$

Since $\Phi(f,x,y)$ is invariant under permutations of x_i and y_i. we can write,

$$\mathbb{P}\Big(\exists f \ S_n f \geq \Phi(f,x,y) + \sqrt{\frac{Wt}{n}}\Big) = \mathbb{P}\Big(\exists f \ R_n f \geq \Phi(f,x,y) + \sqrt{\frac{Wt}{n}}\Big)$$
$$= \mathbb{E}\mathbb{P}_\varepsilon\Big(\exists f \ R_n f \geq \Phi(f,x,y) + \sqrt{\frac{Wt}{n}}\Big). \tag{3.2}$$

For a fixed (x,y) consider a set

$$\mathcal{A} = \{\varepsilon : \sup_{f\in\mathcal{F}}(R_n f - \Phi(f,x,y)) \leq 0\}.$$

By condition (3.1), $\mathbb{P}_\varepsilon(\mathcal{A}) \geq \beta$. If we denote $\mathcal{A}_t = \{\varepsilon : f_c^2(\mathcal{A},\varepsilon) \leq t\}$ then (1.1) implies that

$$\mathbb{P}_\varepsilon(\mathcal{A}_t) \geq 1 - \beta^{-\alpha}\exp\Big(-\frac{\alpha}{\alpha+1}t\Big).$$

Let us take $\varepsilon \in \mathcal{A}_t$ and $\varepsilon' \in \mathcal{A}$. The definition of $\mathcal{A}$ implies that for any $f \in \mathcal{F}$

$$\frac{1}{n}\sum_{i=1}^n \varepsilon_i'(f(y_i) - f(x_i)) \leq \Phi(f,x,y),$$

and, therefore,

$$\frac{1}{n}\sum_{i=1}^n \varepsilon_i(f(y_i)-f(x_i)) - \Phi(f,x,y) \leq \frac{1}{n}\sum_{i=1}^n(\varepsilon_i - \varepsilon_i')(f(y_i)-f(x_i))$$
$$\leq \frac{2}{n}\sum_{i=1}^n |f(y_i)-f(x_i)| I(\varepsilon_i' \neq \varepsilon_i).$$

But since $\varepsilon \in \mathcal{A}_t$, (1.2) implies that one can choose $\varepsilon' \in \mathcal{A}$ so that

$$\frac{2}{n}\sum_{i=1}^n |f(y_i)-f(x_i)| I(\varepsilon_i' \neq \varepsilon_i) \leq \Big(t\frac{4}{n^2}\sum_{i=1}^n (f(y_i)-f(x_i))^2\Big)^{1/2} = \Big(\frac{Wt}{n}\Big)^{1/2}.$$

This proves the theorem.

□

Let us consider a special case of $\Phi(f,x,y)$, which satisfies condition (3.1). Let us note here that application of Talagrand's concentration inequality for two point space as it was implemented in Theorem 2 is not crucial for the examples of this section. It is well known

fact that the chaining technique that we will only use here to bound the $(1-\beta)$-quantile implies tail estimates as well. But it is hard to argue with the fact that the application of Talagrand's inequality even for these examples is more elegant as it immediately provides the tail estimates once the bound for the quantile is obtained.

We will assume from now on that $0 \equiv f \in \mathcal{F}$. Let d be a metric on $\mathcal{F}$. Given $u > 0$ we say that a subset $\mathcal{F}' \subset \mathcal{F}$ is $u-$separated if for any $f \neq g \in \mathcal{F}'$ we have $d(f,g) > u$. Let a *packing number* $D(\mathcal{F}, u, d)$ be the maximal cardinality of a $u-$separated set.

We define

$$\Phi(f,x,y) = Kn^{-1/2} \int_0^{\sqrt{W}/2} (\log D(\mathcal{F}, u, d_{x,y}))^{1/2} du,$$

where

$$d_{x,y}(f,g) = \Big(\frac{1}{n}\sum_{i=1}^n (f(y_i) - f(x_i) - g(y_i) + g(x_i))^2\Big)^{1/2}$$

and $K = K(\beta)$ depends only on β. For example, if $K(\beta) = 8(p+2)^{1/2}$, where p is such that $\sum_{j=2}^\infty j^{-p} < 1-\beta$, then the following theorem holds.

Theorem 3 *If $K(\beta)$ is defined as above then (3.1) holds.*

Proof. The proof is based on standard chaining technique. Let us fix (x,y). Define

$$F = \{(f(y_1) - f(x_1), \ldots, f(y_n) - f(x_n)) : f \in \mathcal{F}\}$$

and

$$d(f,g) = \Big(\frac{1}{n}\sum_{i=1}^n (f_i - g_i)^2\Big)^{1/2}, \quad f, g \in F.$$

Then, if

$$\Phi(f) = K(\beta) n^{-1/2} \int_0^{d(f,0)} (\log D(F, u, d))^{1/2} du,$$

we need to prove that

$$\mathbb{P}_\varepsilon\Big(\sup_{f\in F}\Big(\frac{1}{n}\sum_{i=1}^n \varepsilon_i f_i - \Phi(f)\Big) > 0\Big) < 1-\beta.$$

Let j_0 be defined as

$$j_0 = \inf\{j : D(F, 2^{-j}, d) \geq 2\}.$$

Consider an increasing sequence of sets

$$\{0\} = F_{-\infty} = \ldots = F_{j_0-1} \subseteq F_{j_0} \subseteq F_{j_0+1} \subseteq \ldots$$

such that for any $g \neq h \in F_j$, $d(g,h) > 2^{-j}$ and for all $f \in F$ there exists $g \in F_j$ such that $d(f,g) \leq 2^{-j}$. The cardinality of F_j can be bounded by

$$|F_j| \leq D(F, 2^{-j}, d).$$

For simplicity of notations we will write $D(u) := D(F, u, d)$. If $D(2^{-j}) = D(2^{-j-1})$ then in the construction of the sequence (F_j) we will set F_j equal to F_{j+1}. We will now define the sequence of projections $\pi_j : F \to F_j$, $j \geq 0$ in the following way. If $f \in F$ is such that $d(f, 0) \in (2^{-j-1}, 2^{-j}]$ then set $\pi_0(f) = \ldots = \pi_j(f) = 0$ and for $k \geq j+1$ choose $\pi_k(f) \in F_k$ such that $d(f, \pi_k(f)) \leq 2^{-k}$. In the case when $F_k = F_{k+1}$ we will choose $\pi_k(f) = \pi_{k+1}(f)$. This construction implies that $d(\pi_{k-1}(f), \pi_k(f)) \leq 2^{-k+2}$. Let us introduce a sequence of sets

$$\Delta_j = \{g - h : g \in F_j, h \in F_{j-1}, d(g, h) \leq 2^{-j+2}\}, \quad j \geq j_0,$$

and let $\Delta_j = \{0\}$ if $D(2^{-j}) = D(2^{-j+1})$. The cardinality of Δ_j does not exceed

$$|\Delta_j| \leq |F_j|^2 \leq D(2^{-j})^2.$$

By construction any $f \in F$ can be represented as a sum of elements from Δ_j

$$f = \sum_{j \geq j_0} (\pi_j(f) - \pi_{j-1}(f)), \quad \pi_j(f) - \pi_{j-1}(f) \in \Delta_j.$$

Let

$$I_j = n^{-1/2} \int\limits_{2^{-j-1}}^{2^{-j}} (\log D(u))^{1/2} du$$

and define the event

$$A = \bigcup_{j=j_0}^{\infty} \Big\{ \sup_{f \in \Delta_j} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f_i \geq K I_j \Big\}.$$

On the complement A^c of the event A we have for any $f \in F$ such that $d(f, 0) \in (2^{-j-1}, 2^{-j}]$

$$\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f_i = \frac{1}{n} \sum_{k \geq j+1} \sum_{i=1}^{n} \varepsilon_i (\pi_k(f) - \pi_{k-1}(f))_i \leq \sum_{k \geq j+1} K I_k$$
$$\leq K n^{-1/2} \int\limits_{0}^{2^{-j-1}} (\log D(u))^{1/2} du \leq K n^{-1/2} \int\limits_{0}^{d(f,0)} (\log D(u))^{1/2} du.$$

It remains to prove that for some constant $K(\beta)$, $P(A) < 1 - \beta$. Indeed,

$$P(A) \leq \sum_{j=j_0}^{\infty} P\Big(\sup_{f \in \Delta_j} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f_i \geq K I_j \Big)$$
$$\leq \sum_{j=j_0}^{\infty} |\Delta_j| \exp\Big\{ -\frac{n K^2 I_j^2}{2^{-2j+4}} \Big\} I\big(D(2^{-j}) > D(2^{-j+1}) \big)$$
$$\leq \sum_{j=j_0}^{\infty} \exp\Big\{ 2 \log D(2^{-j}) - \frac{n K^2 I_j^2}{2^{-2j+4}} \Big\} I\big(D(2^{-j}) > D(2^{-j+1}) \big),$$

since for $f \in \Delta_j$ $n^{-1}\sum_{i=1}^n f_i^2 \le 2^{-2j+4}$. The fact that $D(u)$ is decreasing implies

$$\frac{n^{1/2} I_j}{2^{-(j+1)}} \ge (\log D(2^{-j}))^{1/2}$$

and, therefore,

$$P(A) \le \sum_{j=j_0}^{\infty} \exp\{-\log D(2^{-j})(K^2 2^{-6} - 2)\} I\big(D(2^{-j}) > D(2^{-j+1})\big)$$
$$\le \sum_{j=j_0}^{\infty} \frac{1}{D(2^{-j})^p} I\big(D(2^{-j}) > D(2^{-j+1})\big) \le \sum_{j=2}^{\infty} \frac{1}{j^p} < 1 - \beta,$$

for $p = K(\beta)^2 2^{-6} - 2$ big enough. We used the fact that $D(2^{-j_0}) \ge 2$.

□

Example (Uniform entropy conditions). Let us introduce a uniform packing numbers $D(\mathcal{F}, u)$ as any function such that

$$\sup_Q D(\mathcal{F}, u, L_2(Q)) \le D(\mathcal{F}, u)$$

where the supremum is taken over all discrete probability measures. One can easily check that

$$\Big(\frac{1}{n}\sum_{i=1}^n (f(x_i) - f(y_i) - g(x_i) + g(y_i))^2\Big)^{1/2} \le 2\Big(\frac{1}{2n}\sum_{i=1}^n \big((f(x_i) - g(x_i))^2 + (f(y_i) - g(y_i))^2\big)\Big)^{1/2}$$

and, therefore, in the case when the packing numbers are bounded uniformly we get,

$$D(\mathcal{F}, u, d_{x,y}) \le D(\mathcal{F}, u/2).$$

Hence,

$$\mathbb{E}_y \Phi(f, x, y) \le K(\beta) n^{-1/2} \mathbb{E}_y \int_0^{\sqrt{W}/2} (\log D(\mathcal{F}, u/2))^{1/2} du$$
$$\le 2K(\beta) n^{-1/2} \int_0^{\sqrt{V}/4} (\log D(\mathcal{F}, u))^{1/2} du.$$

Corollary 3 *For any* $t \ge \log \beta^{-1}$,

$$\mathbb{P}\Big(\exists f \in \mathcal{F}\ Q_n f \ge \frac{2K(\beta)}{n^{1/2}} \int_0^{\sqrt{V}/4} (\log D(\mathcal{F}, u))^{1/2} du + \sqrt{\frac{Vt}{n}}\Big) \le \exp(1 - (\sqrt{t} - \sqrt{\log \beta^{-1}})^2).$$

□

In the case of VC-subgraph classes with VC dimension d (for definition, see van der Vaart and Wellner (1996)), the result of [5] gives

$$D(\mathcal{F}, u) \le e(d+1)\Big(\frac{2e}{u^2}\Big)^d,$$

and, therefore, the following corollary.

Corollary 4 *(Normalization by variance). There exists K that depends only on β such that for any $t \geq \log \beta^{-1}$,*

$$\mathbb{P}\Big(\exists f \in \mathcal{F}\ \frac{Q_n f}{\sqrt{V}} \geq K\sqrt{\frac{d\log n}{n}} + \sqrt{\frac{t}{n}}\Big) \leq \exp(1 - (\sqrt{t} - \sqrt{\log \beta^{-1}})^2).$$

□

Let us rewrite V as

$$V = V(x) = 4(\mathrm{Var} f + \mathrm{Var}_n f + (Pf - P_n f)^2) = 4(\mathrm{Var} f + \mathrm{Var}_n f + (Q_n f)^2),$$

where

$$\mathrm{Var}_n f = \frac{1}{n}\sum_{i=1}^n (P_n f - f(x_i))^2$$

is a sample variance. If we denote

$$U = K\sqrt{\frac{d\log n}{n}} + \sqrt{\frac{t}{n}},$$

then one can solve the inequality of Corollary 4 for $Q_n f$ to get

$$\mathbb{P}\Big(\exists f \in \mathcal{F}\ |Q_n f| \geq 2U\Big(\frac{\mathrm{Var} f + \mathrm{Var}_n f}{1 - 4U^2}\Big)^{1/2}\Big) \leq 2\exp(1 - (\sqrt{t} - \sqrt{\log \beta^{-1}})^2).$$

Let us compare this to an "optimistic" inequality of Vapnik and Chernonenkis [18], which states that if $\mathcal{F} = \{f : \Omega \to \{0,1\}\}$ is a VC-class of indicator functions with VC dimension d, then with probability at least $1 - e^{-t/4}$, for all $f \in \mathcal{F}$

$$\frac{1}{n(Pf)^{1/2}}\sum_{i=1}^n (Pf - f(x_i)) \leq 2\Big(\frac{d}{n}\log\frac{2en}{d} + \frac{t}{n}\Big)^{1/2}.$$

Compared to the inequality of Vapnik and Chervonenkis our inequality controls the deviation of $P_n f$ from Pf in both directions, no assumptions are made on the boundedness of functions $f \in \mathcal{F}$, and the deviation is controled by the mixture of variance and sample variance rather than by expectation Pf, which can be considered as a significant improvement.

Example (The case of one function). When $\mathcal{F}$ consists of one function f we will simply write $f(X) = \xi$. Let us take $\beta = 1/2$ and let

$$\Phi(\xi) = \mathbb{E}_{\xi'} M_\varepsilon\Big(\frac{1}{n}\sum_{i=1}^n \varepsilon_i(\xi_i - \xi_i')\Big) = 0.$$

Obviously, with this choice of β and Φ condition (3.1) holds and Theorem 2 implies

$$\mathbb{P}\Big(|\bar{\xi} - \mathbb{E}\xi| \geq 2\Big(\frac{(\mathrm{Var}\xi + \mathrm{Var}_n\xi + (\mathbb{E}\xi - \bar{\xi})^2)t}{n}\Big)^{1/2}\Big) \leq 2\exp\Big(1 - (\sqrt{t} - \sqrt{\log 2})^2\Big).$$

Solving the inequality for $|\bar{\xi} - \mathbb{E}\xi|$ we get

$$\mathbb{P}\Big(|\bar{\xi} - \mathbb{E}\xi| \geq 2\Big(\frac{(\mathrm{Var}\xi + \mathrm{Var}_n\xi)t}{n - 4t}\Big)^{1/2}\Big) \leq 2\exp\Big(1 - (\sqrt{t} - \sqrt{\log 2})^2\Big). \tag{3.3}$$

One should compare this to Bernstein type inequalities. First of all, we don't assume any moment conditions other than the existance of variance of ξ. Second, in Bernstein's inequality

$$|\bar{\xi} - \mathbb{E}\xi| \lesssim \Big(\frac{t\mathrm{Var}\ \xi}{n}\Big)^{1/2} \text{ for } t \leq n\mathrm{Var}\ \xi,$$

$$|\bar{\xi} - \mathbb{E}\xi| \lesssim \frac{t}{n} \text{ for } t \geq n\mathrm{Var}\ \xi,$$

whereas (3.3) gives

$$|\bar{\xi} - \mathbb{E}\xi| \leq 2\Big(\frac{2(\mathrm{Var}\xi + \mathrm{Var}_n\xi)t}{n}\Big)^{1/2} \text{ for } t \leq n/8.$$

This, basically, means that the deviation of the average $\bar{\xi}$ from the expectation $\mathbb{E}\xi$ can be large only when the sample variance is large.

Acknowledgment. We want to thank Michel Talagrand for some valuable comments and suggestions.

References

[1] Boucheron, S., Lugosi, G., Massart, P. (2000) A sharp concentration inequality with applications. *Random Structures Algorithms* **16** 277 - 292.

[2] Dembo, A. (1997) Information inequalities and concentration of measure. *Ann. Probab.* **25** 527 - 539.

[3] Dudley, R.M. (1999) Uniform Central Limit Theorems. Cambridge University Press.

[4] Giné, E., Götze, F., Mason, D. (1997) When is the Student t-statistics asymptotically standard normal? *Ann. Probab.* **26** 1514 - 1431.

[5] Haussler, D. (1995) Sphere packing numbers for subsets of the boolean $n-$cube with bounded Vapnik-Chervonenkis dimension. *J. Combin. Theory Ser. A* **69** 217 - 232.

[6] Ledoux, M. (1996) On Talagrand's deviation inequalities for product measures. *ESAIM: Probab. Statist.* **1** 63 - 87.

[7] Ledoux, M. and Talagrand, M. (1991) Probability in Banach Spaces. Springer-Verlag, New York.

[8] Massart, P. (2000) About the constants in Talagrand's concentration inequalities for empirical processes. *Ann. Probab.* **28** 863 - 885.

[9] Panchenko, D. (2001) A note on Talagrand's concentration inequality. *Elect. Comm. in Probab.* **6** 55 - 65.

[10] Panchenko, D. (2002) Some extensions of an inequality of Vapnik and Chervonenkis. *Elect. Comm. in Probab.* **7**

[11] Rio E. (2000) Inégalités exponentielles pour les processus empiriques. *C.R. Acad. Sci. Paris,* t.330, Série I 597-600.

[12] Rio E. (2001) Inégalités de concentration pour les processus empiriques de classes de parties. *Probab. Theory Relat. Fields* **119** 163-175.

[13] Shao, Q.-M. (1997) Self-normalized large deviations. *Ann. Probab.* **25** 285 - 329.

[14] Talagrand, M. (1995) Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de l'I.H.E.S.* **81** 73-205.

[15] Talagrand, M. (1996) New concentration inequalities in product spaces. *Invent. Math.* **126** 505-563.

[16] van der Vaart, A., Wellner, J. (1996) Weak Convergence and Empirical Processes: With Applications to Statistics. John Wiley & Sons, New York.

[17] Vapnik, V.N., Chervonenkis, A.Ya. (1968) On the uniform convergence of relative frequencies of event to their probabilities. *Soviet Math. Dokl.***9** 915 - 918.

[18] Vapnik, V.N. (1998) Statistical Learning Theory. Wiley, New York.

Department of Mathematics and Statistics
The University of New Mexico
Albuquerque, NM 87131–1141
e-mail: panchenk@math.unm.edu